Intronic AT skew is a defendable proxy for germ-line transcription but does not predict crossing-over or protein evolution rates in *Drosophila melanogaster*

Claudia C. Weber and Laurence D. Hurst*

Department of Biology and Biochemistry, University of Bath, Bath, BA2 7AY, UK

*email: **l.d.hurst@bath.ac.uk**

Recent evidence suggests that germ-line transcription may affect both protein evolutionary rates, possibly mediated by repair processes, and recombination rates, possibly mediated by chromatin and epigenetic modification. Here we test these propositions in *Drosophila melanogaster.* The challenge for such analyses is to provide defendable measures of germline gene expression. Intronic AT skew is a good candidate measure as it is thought to be a consequence, at least in part, of transcription coupled repair. Prior evidence suggests that intronic AT skew in *D. melanogaster* is not affected by proximity to intron extremities and differs between transcribed DNA and flanking sequence. We now also establish that intronic AT skew is a defendable measure of germline expression as a) it is more similar than expected by chance between introns of the same gene (which is not accounted for by physical proximity), b) is correlated with male germ-line expression, and c) is more pronounced in broadly expressed genes. Furthermore, d) a trend for intronic skew to differ between 3' and 5' ends of genes is particular to broadly expressed genes. Finally, e) controlling for physical distance, introns of proximate genes are most different in skew if they have different tissue specificity. We find that intronic AT skew, employed as a proxy for germ-line transcription, correlates neither with recombination rates nor with the rate of protein evolution. We conclude that there is no *prima facie* evidence that germ-line expression modulates recombination rates or monotonically affects protein evolution rates in *D. melanogaster*.

Key words: Drosophila melanogaster; germ-line expression; somatic gene expression; tissue specificity; recombination; nucleotide asymmetry; intronic AT skew.

There is increasing interest in the possibility that many features of gene evolution may be a direct consequence of processes operating in the germline, rather than the action of selection (Duret and Galtier 2009). Biased gene conversion during recombination, for example, is thought to explain much of the between-gene variation in GC content at sites under weak selection (Duret and Galtier 2009; Eyre-Walker and Hurst 2001) and may also explain why some genes appear to be under positive selection (e.g. show high Ka/Ks ratios) in the absence of positive selection (Berglund et al. 2009; Galtier et al. 2009; Galtier and Duret 2007; Nagylaki 1983). Here we examine a further possibility, namely that transcription in the germline might directly affect protein evolution and recombination rates mediated by, amongst other things, effects on chromatin, epigenetic marks and involvement with repair processes.

Preliminary evidence from a variety of taxa suggests the possibility that germline transcription, or modifiers of transcription, may impact on protein evolution (Weber and Hurst 2009) and recombination rates (Necsulea et al. 2009b; Sigurdsson et al. 2009). We recently found that in the pre-meiotic period in yeast genes that are highly expressed tend to be prone to non-meiotic double-strand breaks, possibly a consequence of the repair of stalled transcriptional forks. These same genes have low rates of protein evolution, even controlling for subsequent rates of crossing over (Weber and Hurst 2009). This low rate of evolution could, we suggest, plausibly be a direct consequence of repair processes associated with resolution of double strand breaks.

A possible connection between germline transcription and recombination has long been considered. An intuitive assumption has been that transcriptionally active, open chromatin should be more prone to recombination (see e.g. Borde et al. 2009; Mancera et al. 2008; Wu and Lichten 1994) implying a positive association with transcription. Consistent with this we found that, in yeast, genes that are transcriptionally active premeiotically tend to have high recombination rates (Weber and Hurst 2009). Recombination hotspots can, however, occur outside of open chromatin (Berchowitz et al. 2009). Moreover in mammals genes expressed in many tissues tend to have low recombination rates (Necsulea et al. 2009b). Similarly, confirming that the above correlation is in part reflective of the notion that genes expressed in many tissues are also germline expressed, it has recently been reported (using human germline specific expression data), that germline expressed genes have low recombination rates (McVicker and Green 2010). This runs counter to the notion that transcriptionally active, open chromatin domains should be those that recombine most. While the simple notion that open chromatin is recombinogenic may be naïve, a connection of some form between chromatin status and recombination is consistent with the observation that, controlling for expression breadth, both imprinted and monoallelically expressed genes show high recombination rates (Necsulea et al. 2009b), a thesis given weight by the discovery of the relationship between histone H3 lysine 4 trimethylation and recombination in *S. cerevisiae* (Borde et al. 2009) and mice (Buard et al. 2009) and by a correlation between methylation and recombination (Sigurdsson et al. 2009).

Here we ask whether germline transcription may be coupled with protein evolution and recombination rates in *Drosophila*. *Prima facie* a possible connection between transcription and recombination may be expected here as, just as in mammals, prior

reports suggest that tissue specific genes have higher recombination rates (see Table S2 in Larracuente et al. 2008). Similarly, genes expressed in many tissues tend to be slow evolving (Larracuente et al. 2008), as is also observed in mammals (Liao et al. 2006). This latter result is typically considered to reflect a greater opportunity for efficient purifying selection in broadly expressed genes. If, however, as often assumed (see e.g. Duret and Mouchiroud 2000; Majewski 2003; Webster et al. 2004), broadly expressed genes are more likely to be germline expressed genes, then we would need to discount any direct effect of germline transcription. Similarly, protein rates of evolution are known to correlate with the local recombination rate, a feature considered supportive of Hill-Robertson interference (see e.g Bachtrog 2003; Betancourt and Presgraves 2002; Betancourt et al. 2009; Haddrill et al. 2008; Presgraves 2005; Zhang and Parsch 2005). To clearly interpret any such correlation, however, it is important to be able to exclude direct effects of germline expression on either parameter.

In multicellular organisms, addressing these questions is problematic owing to difficulties measuring gene expression at all time points through germline development. To overcome this issue, somatic expression patterns are commonly used to extrapolate germline expression, under the premise that a gene that is commonly expressed in soma is more likely to be expressed in germline (see e.g. Duret and Mouchiroud 2000; Majewski 2003; Webster et al. 2004). Such an extrapolation is of no use for our purposes. As both highly and broadly expressed genes have low rates of protein evolution (Larracuente et al. 2008; Pal et al. 2001), probably in large part as a direct consequence of selection mediated by expression level, in order to ask whether germline transcription *per se* affects protein evolutionary rates, we require a measure of germline expression that is not extrapolated from somatic patterns. To this end we use

intronic AT skew as this most likely reflects mutational biases resulting from germ-line transcription.

Intronic AT skew is thought partly to be a result of transcription-coupled repair or other germline transcriptional processes (see Comeron 2004; Green et al. 2003; Majewski 2003; Mugal et al. 2010; Svejstrup 2002; Touchon et al. 2004). Importantly, when averaged across 1kb windows the skew is different in transcribed sequence compared to flanking non-transcribed sequence, the transition being observed immediately 5' of the transcription start site (Touchon et al. 2004). In *D. melanogaster*'s introns, T is enriched over A compared with the flanking non-coding sequences, suggesting that germline expression skews towards increasing T content. Unlike intronic GC skew, intronic AT skew is not more pronounced near intron extremities, suggesting that it is not affected by splice related constraints in *D. melanogaster* (Touchon et al. 2004). Note that this behaviour of intronic AT skew is peculiar to Drosophila (Touchon et al. 2004), a feature that motivates the choice of study species.

Before concluding that intronic AT skew is a defendable measure of germline expression it is necessary to test further predictions of the transcription model. We start therefore by gauging skew against expected benchmarks. Notably, we should expect that intronic AT is more similar than expected by chance between introns from the same gene. We should also expect it to correlate with breadth of expression, assuming breadth tells us something about germline expression. Limited expression data from male germline is now available and, naturally, we should see intronic skew correlating with this. Note that we do not postulate that intronic AT skew is solely

owing to germline transcription. Replication is, for example, also likely to be a cause of bias (see e.g Touchon et al. 2005). Rather we ask whether a component of AT skew is parsimoniously attributable to germline transcription, which if the case permits us to use it as a proxy, albeit not necessarily a perfect proxy.

Methods:

In order to prevent unnecessary sample size reductions, all data were mapped to primary FlyBase identifiers, as some tables contained secondary identifiers, FlyBase identifiers or CG numbers. Where a given identifier was represented more than once, values were averaged.

Estimates of recombination rate

Estimates of recombination rate per gene were obtained from Larracuente et al. (2008). Their set is derived from Release 4.3 of the *D. melanogaster* genome and includes two methods of approximating the crossover rate: ACE and RP. ACE is based on local estimates of the relationship between genetic and physical map positions across polytene bands, whereas RP is the slope of the third order regression polynomial at the midpoint of each gene (see Hey and Kliman 2002). Both rates correlate (rho = 0.9631, $p < 2.2 \times 10^{-16}$ for RP; rho = 0.6898, $p < 2.2 \times 10^{-16}$ for ACE) with those measured by Hey and Kliman (2002), although the new data are expected to be more accurate on account of the availability of a greater number of genes with known genetic and physical map locations.

Although it has been argued that smoothed data might give a better reflection of average long-term recombination rates over time (Marais et al. 2003), data from *Drosophila pseudoobscura* suggests that fine-scale recombination rates are more

strongly correlated with diversity and interspecies difference than are broad-scale estimates (Kulathinal et al. 2008). There is also evidence for considerable heterogeneity in *D. melanogaster* recombination rates (Singh et al. 2009). Although whole-genome recombination estimates are presently limited in resolution, it is nevertheless important to consider the possible drawbacks and advantages of different methods of estimating recombination. While the polynomial approach has a smoothing effect that may obscure regional variation, the sliding window approach is more susceptible to noise due to errors in genetic and physical maps (Marais et al. 2003). Therefore, both ACE and RP were considered throughout to ensure relationships are not artefacts of a particular method.

Expression

Mean whole fly expression

The mean of male and female whole fly expression for *D. melanogaster* was calculated for each gene from ftp://ftp.ncbi.nih.gov/pub/geo/DATA/SeriesMatrix/GSE6640/GSE6640-GPL4629_series_matrix.txt.gz (Zhang et al. 2007).

Tissue Bias

Adult tissue expression data from Fly Atlas (Chintapalli et al. 2007) was used to calculate τ (see Liao et al. 2006; Yanai et al. 2005).

(1)

$$\tau = \frac{\sum_{j=1}^{n}\left(1 - \frac{logS(j)}{logS_{max}}\right)}{n-1}$$

For genes where expression was detected on only one array for a given tissue, *logS(j)* for that tissue was set to 0. All genes where all tissues were hence set to 0 were excluded from the analysis. Logging the ratio of tissue expression to maximal gene expression reduces the influence of large differences between maximal gene expression and tissue expression. Tau takes into account two things: a) How many tissues is expression detected in? b) How does the level of expression in each tissue compare to maximal gene expression?

Nucleotide skew

Germline expression is thought to lead to compositional asymmetries in transcribed sequences (Touchon et al. 2004). Intron sequences and fourfold degenerate third sites from *D. melanogaster* coding sequences were obtained from FlyBase (release 5.22). Only introns at least 160bp long were considered and 30bp were trimmed from each end. For both introns and fourfold degenerate sites in coding sequences, A-T/A+T and G-C/G+C were calculated to give a measure of nucleotide skew.

Testis expression

D. melanogaster expression data from mitotic, meiotic and postmeiotic stages of spermatogenesis were obtained from SpermPress (Vibranovski et al. 2009).

Divergence

ω (Ka/Ks) values were taken from ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/paml/. Rows where ω = 999 (indicating a PAML error) were removed. These data are based on six of the twelve available species only because synonymous site distances are too great to give accurate d_S and ω estimates (see Haddrill and Charlesworth 2008; Larracuente et al.

2008). *D. melanogaster* branch model distances were used, as the available recombination rates and expression measures are also from *D. melanogaster*. This accounts for the possibility that recombination hotspots are prone to moving over time between *Drosophila* species, for which some evidence exists: For instance, X-telomeric recombination rates in *D. melanogaster* are reduced about 20-fold compared to *D. yakuba* (Takano-Shimizu 2001).

Statistics

All statistical analyses were performed in R. Some complex relationships might not be adequately captured if only linear associations are tested for. Therefore Spearman's rho, hereafter referred to as rho, which assumes monotonicity but not linearity, was used.

Results

Intronic AT skew is similar between introns from the same gene

Having ruled out using intronic GC skew due to the strong influence of cytosine enrichment on splice boundaries (Touchon et al. 2004), we ask whether AT skew might be being influenced by germline expression. If AT skew is associated with germline transcription then we should expect it to be fairly consistent within a gene. To ask whether genes have prototypical AT skews we consider the median difference in AT skew between introns from the same gene and compare this to a randomized null, where we preserve the intron structure of each gene but randomize observed skew values.

For genes with two or more introns all possible pairwise distances between introns of the same gene were calculated. Prior to calculating the differences, genes were sorted by the start positions of their first introns. Where genes with overlapping coordinates were found, the gene overlapping its left neighbour (if that neighbour had not yet been removed itself) was deleted. Because genes were sorted by starts, not midpoints, this ought not to introduce a bias towards preserving shorter genes. Likewise, overlapping intron sequences were removed. Intron skews were then randomised 10000 times, i.e. all intron pairs were preserved but skew values were shuffled so that each intron was assigned a random skew, and pairwise distances were again calculated for each randomization. For AT skew the observed difference in skew between introns of the same gene was smaller than any randomized distance ($p < 10^{-5}$, observed median distance = 0.0768, median of randomizations = 0.0823; see Figure 1).

To determine whether greater similarity between introns of the same gene is due to physical proximity, we considered the difference in skew between introns whose midpoints were located between 800 and 1200bp apart for both introns of the same and randomly selected introns of different genes (1000 randomly selected pairs per chromosome arm). As introns from different genes on chromosome 4 did not meet the distance criterion, only the euchromatic regions of the remaining chromosome arms were included, and pairs with missing tissue specificity information were filtered out. As expected if germline transcription affects AT skew, the difference between intron pairs of the same gene (median distance = 0.0967, n = 1140) was significantly smaller than for intron pairs of different genes (median distance = 0.1214, n = 3773; Wilcoxon test p = 0.0014). When intron pairs from different genes located 800-1200bp apart that

are also dissimilar in terms of tissue specificity, i.e. where one gene is in the upper and one in the lower quartile for tau, are considered this difference becomes even more profound (median distance = 0.2387, n = 458; Wilcoxon test $p < 2.2 \times 10^{-16}$). Thus, physical distance does not appear to account for the greater similarity in AT skew for introns of the same gene. Meanwhile, introns of genes expected to have different expression patterns in the germline differ even when they are close together. These results are consistent with germline transcription affecting AT skew, above and beyond any replication or genomically regional effects.

We might presume that all introns in the same gene are subject to transcription-related mutational biases at equal rates, while prior evidence (see e.g. Polak and Arndt 2008) has identified mutational biases that differ 5'-3' along a gene, at least in mammals. Many reasons could be given for this. Alternative transcripts differing in the 3' ends but employing the same 5' end can ensure a given 5' base is transcribed more than a 3' base. By contrast, alternative 5' start sites resolving to the same 3' end can mean that terminal exons are transcribed more than 5' exons. Transcription can also prematurely abort. In about 80% of genes in human embryonic stem cells there is an initiation of transcription, but just 50% of the genes are fully transcribed (Guenther et al. 2007). Moreover, flies have unusually high levels of sense-antisense pairs (Sun et al. 2006) and such pairs tend to overlap at 3' ends (Sun et al. 2005). The formation of R loops (single strand DNA-RNA hybrids) preferentially at the 5' ends of transcribed genes has also been suggested to contribute to difference in substitution processes along a gene (Polak and Arndt 2008). This could in principle either increase or decrease skew at the 5' end. Given the above we should expect that first introns should be more similar to second introns than to the last intron. We should also expect to see a change in skew

5'->3' that is most pronounced in broadly expressed genes, these being more likely to be germline expressed.

When we calculate the median distance between the most 5' pairs of introns for each gene, we find that they are also significantly more similar than the randomised comparisons (observed median distance = 0.0752, $p < 10^{-5}$). We also observe, as expected, that the absolute difference in skew between first and second introns is smaller than the absolute difference between first and last introns of genes (median distance for first and last = 0.0824; paired Wilcoxon test for median difference between 1^{st} versus 2^{nd} and 1^{st}-last comparisons -0.0106, $p = 1.348x10^{-7}$, n= 3307). Underpinning the tendency for skew to vary monotonically through a gene, we find no significant difference between first and last introns of the same gene compared to random expectation (observed = 0.0824, expected = 0.0823, p = 0.5437).

More generally, we observe the predicted change in intronic AT skew and intron position relative to the most 5' and 3' introns for a given gene exclusively for putatively germline expressed genes. For this analysis genes with a low tissue specificity value, defined as the lowest quartile for tau, were considered broadly expressed. We then consider the overall correlation between AT skew and relative intron position. This is significant for broadly expressed genes (rho = - 0.1148, $p < 2.168x10^{-7}$, n=2029), but not other genes (rho = -0.0166, p = 0.1962, n = 6084; see Figure 2). The same is true considering absolute intron positions (low tau rho= -0.1046, $p = 2.348x10^{-6}$; others, rho= 0.0018, p= 0.8887). More generally, and in accord with the above, there is a positive relationship between tau and the rank correlation of intron positions within a gene against skew (rho = 0.0620, p = 0.0120;

see Figure 3). These results are consistent with the hypothesis that broadly expressed genes show a greater change in the magnitude of skew from 5' to 3' positions, and hence that such skews reflect germline transcriptional processes.

Intronic AT skew correlates with breadth of expression and male germline expression

If intronic AT skew is related to germline expression, then we might expect such values to correlate with expression measures. Tau, a measure of tissue specificity, is indeed positively correlated with A-T/A+T in all introns > 160bp (rho = 0.1195, $p < 2.2 \times 10^{-16}$, n = 5438; see Figure 4) and long (> 2100bp) introns (rho = 0.1588, $p = 5.088 \times 10^{-14}$, n = 2223). This is consistent with T enrichment of genes with germline expression (low tau). The weakness of GC skew as a measure of germline expression is further corroborated by analysis of correlation with expression measures. We find that tau is not correlated with G-C/G+C for long introns (rho = 0.0045, p = 0.8325, n = 2223) and only weakly so for all introns > 160bp (rho = 0.0477, p = 0.0004, n = 5438).

If intronic AT skew is a suitable proxy for germline expression we might expect it to correlate with direct measures of germline expression. Recently male germline expression data from Vibranovski et al. (2009) has been produced. This data examines transcriptional activity in male germ cells around the time of meiosis. As recombination only occurs in females this data is of limited use in asking whether germline expression interacts with recombination. It also does not measure expression in early germline events nor what happens in female germline, both of which are likely to influence AT skew. Nonetheless, as we would expect AT skew in introns is

negatively correlated with mitotic expression (rho = -0.1418, $p < 2.2 \times 10^{-16}$, n = 6188), meiotic expression (rho = -0.1175, $p < 2.2 \times 10^{-16}$) and postmeiotic expression (rho = -0.0913, $p = 6.185 \times 10^{-13}$). This provides further evidence that intronic AT skew is shaped by germline transcription-coupled processes. That the relationship observed is stronger for intronic AT skew is consistent with GC skew being a weaker predictor of germline transcription. Intronic GC skew is less strongly associated with mitotic (rho = -0.0876, $p = 5.15 \times 10^{-12}$), meiotic (rho = -0.0906, $p = 9.06 \times 10^{-13}$) and postmeiotic male germline expression (rho = -0.0695, $p = 4.345 \times 10^{-8}$).

Replication alone does not account for the association between tissue specificity and AT skew

Nucleotide skews may also arise from biased mutational processes associated differentially with the leading and lagging strands during replication (see e.g. Touchon et al. 2005). If in the germline a given gene is consistently passaged by the replication fork in a particular direction then a net skew could result. Definitively rejecting this possibility without knowledge of origins of replication in germline is difficult. Quite how this model would account for a 5' to 3' decline in skew through a gene being particular to housekeeping genes is not, however, obvious. Moreover, that proximate genes are more dissimilar in skew, controlling for physical distance, when they differ in tissue specificity suggests that replication cannot be the only source of AT skew.

To explain the overall correlation between AT skew and expression breadth such a replication model would require broadly expressed genes to be more likely to co-orient with the replication fork. When replication fork direction is inferred from origins that have not been experimentally verified (Huvet et al. 2007), such an effect has been claimed in mammals. However, where origins have been experimentally verified the

same trends are not observed (Necsulea et al. 2009a). While we do not know the location of origins, for broadly expressed genes to co-orient with the replication fork more commonly than expected, neighbouring pairs of such genes would need to be co-oriented (i.e. on the same strand) more often than expected. When we consider pairs of adjacent genes that both have either a low or high tissue specificity, defined as being in the lower or upper quartile for tau, we find that low-tau pairs are less likely to be co-oriented than expected (713 not co-oriented; 280 co-oriented), whereas high-tau pairs are more likely to be co-oriented (495 not co-oriented, 616 co-oriented; chi sq = 158.1115, df = 1, $p < 2.2x10^{-16}$). This is consistent with the observation that divergently transcribed gene pairs in *D. melanogaster* are mostly housekeeping (i.e. broadly expressed) genes (Yang and Yu 2009).

The above observations suggest that replication does not account for the association between tissue specificity and AT skew. Further, Touchon et al. (2004) argue that, even if transcription and replication were co-oriented, in order to produce the observed sharp transitions in averaged skew at gene extremities, replication would have to terminate abruptly at 3' gene ends. There is no obvious reason why this should be so. Moreover that 5' and 3' introns differ in skew but only in broadly expressed genes is hard to account for in terms of replication related mechanisms.

Additionally, we can perform tests comparable to those above. If nucleotide skews are largely due to germline transcription and not effects of replication, the difference in intronic skew between any given gene and its neighbour should, all things being equal, be no different than the difference between the same gene and a random gene on the same chromosome. By contrast if a gene's skew is owing to a replication fork going

through it more in one direction than the other, then we should also expect neighbouring genes to show similar skew levels, assuming the fork passes through both in the same direction.

In order to determine whether this is the case, the median difference between the weighted (by length) average intron skew of each gene and its neighbour was calculated. If transcription has no effect on skew and if all skew was the result of replication, then which strand a gene resides on is irrelevant. Thus, for genes on opposite strands, the skew of the sequence on the complementary strand was multiplied by -1 to reflect any mutational process on a given DNA strand, rather than reflecting transcriptional orientation. Overlapping genes were removed as above. Genes with no known position in the genome assembly were also excluded from the analysis. Gene order was then randomised 10000 times, treating each chromosome and eu- and heterochromatic sequences separately.

There was no significant difference between the median (weighted by the number of introns in each chromosome class) difference in skew between genes for the randomised data and the median observed in the actual genome for AT (observed median 0.0663, simulated median 0.0674, $p = 0.1255$). This is also the case if only those sets of adjacent genes that are below the third quartile for distance between midpoints (<23.580 kb) are compared to the randomised differences in AT skew (observed median 0.0711, $p = 1$). Hence, the lack of similarity between adjacent genes cannot be explained by the presence of genes that are labelled “adjacent” but that actually reside relatively far apart (and potentially in different replication blocks). However, if we restrict analysis to those gene pairs on the same strand (and hence

putatively affected in the same way by transcription and replication) we see a weak but significant difference (observed median 0.0652, p = 0.0126).

That we see an effect only when the putative transcriptional and a putative replicatory origin of skew both operate in the same direction provides further evidence that transcription contributes to skew. But how can we explain why such genes are weakly similar in their skew? We conjecture that this might reflect local genomic similarity in expression profiles. Note that in the above test our null hypothesis is that neighbouring genes are no more similar in their expression than two random genes. Were this not so then we could expect to see regional effects that would also be consistent with germline expression effects. In mammals house-keeping genes cluster (Lercher et al. 2002). If house-keeping genes are also germline expressed we might expect to see germline expression clusters. Is this seen in flies? We asked whether adjacent genes are more similar to one another in terms of tissue specificity.

The simulated median difference (0.2813, 10000 replicates) is significantly larger than the observed median difference (0.2301, $p < 10^{-5}$). The difference becomes more profound when the upper quartile for distance between genes is removed (median difference = 0.2189) or only the lower quartile for distances is considered (median difference = 0.2010). Therefore, it appears that genes with similar tissue specificities are indeed more likely to be adjacent to each other than expected by chance, just as observed in humans. How this pattern relates to the clustering of genes with similar expression patterns (de Wit et al. 2008; Spellman and Rubin 2002), remains to be resolved. Given the above, is quite possible that local similarities in expression breadth account in part for local similarities in distance in intronic AT skew for adjacent co-

oriented genes. That skew changes from 5' to 3' in broadly expressed, but not other genes, and that breadth of expression is correlated with intronic AT skew is not explained by the replication model. We conclude that intronic AT skew is influenced by germ-line transcription and provides a defendable, albeit potentially weak, proxy for germline expression.

No evidence for a relationship between intronic AT skew and recombination

Assuming that an intronic excess of T over A is a measure of the likelihood of germline gene expression, we can ask whether germline transcription is correlated with recombination rate. We find that neither of the two measures of recombination rate, ACE and RP, are significantly correlated with intronic AT skew and indeed disagree on the sign of any correlation (ACE: rho = -0.0042, p = 0.7423; RP: rho = 0.0150, p = 0.2345, n= 6282). This test might, however, be unfair in that some chromosomal domains, such as centromeres and chromosome 4, are prevented from recombining. Gene expression in these domains cannot thus influence recombination and will add noise to the data. When genes in regions that cannot recombine are excluded (non-recombinogenic domains are defined as having an RP of 0), we recover the same absence of a correlation (for RP, rho = 0.0136, p = 0.3118; for ACE rho = -0.0091, p = 0.4988, n = 5538). If AT skew is an efficient measure of germ-line expression, we thus conclude that we see no effect of germline expression on recombination rates.

No evidence for a relationship between intronic AT skew and sequence divergence.

In yeast we observed that genes prone to premeiotic double strand breaks tended to have low rates of protein evolution, regardless of their crossing-over rate (Weber and Hurst 2009). Such pre-meiotic DSBs may well reflect the rescue of transcriptional forks and, as expected, such genes tended to be highly expressed at the appropriate time. Is germline expression, measured via intronic AT skew, a predictor of rate of protein evolution? In contrast to what we see in yeast, we observe no relationship between intronic AT skew and protein rates of evolution (rho = 0.0044, p = 0.7728, n = 4327), for which we see, as generally observed, a profound correlation with somatic expression profiles (rho = -0.2461, $p < 2.2 \times 10^{-16}$, n = 8386). Similarly, when we employ direct measures of male germline expression and average across the premeiotic, meiotic and postmeiotic stages, we obtain the same lack of a simple association between divergence and testis expression (rho = -0.0032, p = 0.7699, n = 8215). This suggests that protein divergence is, if at all, little affected by transcription-associated processes in germline.

The above analysis, however, masks a more complex picture. For both mitotic (rho = -0.0334, p = 0.0025) and postmeiotic (rho = -0.0488, $p = 9.486 \times 10^{-6}$) testis expression highly expressed genes are significantly more slow evolving. This is comparable to the common finding that highly expressed genes tend to be slow evolving (see e.g. Larracuente et al. 2008). However, unexpectedly, there is a positive relationship between meiotic testis expression and divergence (rho = 0.0246, p = 0.0256). Why we see this is unclear. It may reflect rapid evolution of male-biased genes (Zhang et al. 2004). As the sign of the correlation changes between stages it appears unlikely that some bias associated with transcription importantly affects rates of protein evolution.

For this to be the case transcription and its associated mutational biases would also have to change by stage, for which we are aware of no evidence.

Multiple expression-related factors affect nucleotide skew at synonymous sites

Above we have established that germline transcription coupled repair (or related processes) influence intronic AT skew, but that germline transcription appears not to have any overall impact on recombination or protein evolutionary rates. Might germline transcription then influence anything other than intronic AT skew? Does it for instance influence nucleotide skew at four-fold degenerate sites in exons?

As expected if transcription affects skew at synonymous sites in exons and consistent with a negative relationship between intronic AT skew and testis expression, we find that AT4 skew correlates negatively with mitotic (rho = -0.1953, $p < 2.2 \times 10^{-16}$) meiotic (rho = -0.1806, $p < 2.2 \times 10^{-16}$) and postmeiotic (rho = -0.1917, $p < 2.2 \times 10^{-16}$, n = 6188) testis expression. Similarly, as expected, intronic AT skew is positively correlated with that at exonic four-fold degenerate sites in the same gene (rho = 0.0304, p = 0.0137, n = 6564).

While the above findings are all consistent with the notion that germ-line transcription affects AT skew at synonymous sites, we note one oddity, namely that the correlation between intronic AT skew and AT4 skew in the same gene (above) is strikingly weak. This, we suggest, is because AT skew at synonymous sites is affected by at least two other forces. First, the need to specify exonic splice enhancers (ESEs) near exon ends is expected to greatly skew usage. In humans, ESE motifs are very A rich (49% of

nucleotides) and T poor (12.5%) (Parmley et al. 2007) and the same is likely true in flies (Warnecke and Hurst 2007; Warnecke et al. 2008). Second, selection for translational accuracy affects codon usage and is likely to affect AT skews, as T is favoured over A in the 10% of genes with the highest expression bias (Powell and Moriyama 1997) and relative synonymous codon usage is higher for T-ending codons in highly expressed genes (Duret and Mouchiroud 1999). Note then that these two forces are antagonistic as to their net effect on AT skew.

In accord with the former possibility, when fourfold degenerate AT skew is plotted against distance from intron-exon boundaries for the first 900 bp of coding sequence at each exonic end, a robust negative relationship is observed (rho = -0.3725, p = 3.935 $x10^{-11}$, n = 299, see Figure 5), consistent with the notion that exonic splice enhancers are enriched at exonic ends and reach well into the coding sequence (see also Warnecke and Hurst 2007). Note too that this trend is counter to that expected were we to see a preference for optimal codons closer to the exon boundary (see also Warnecke and Hurst 2007). That fourfold degenerate AT skew is negatively correlated with somatic expression levels (rho = -0.0222, p = 0.0116, n = 12928) is consistent with selection favouring T (above A) ending codons in highly expressed genes.

Intronic GC skew is also thought to be affected by germline transcription, but is a less suitable measure as roughly 1 kb of intron sequence adjacent to splice sites is known to be enriched for cytosine in *D. melanogaster* (Touchon et al. 2004), probably to aid splice site recognition. In accord with the notion that intronic GC skew is not simply reflecting germline transcription (and neither might GC skew at four-fold degenerate sites), GC skew in introns is not correlated to that seen at four-fold sites (rho = -0.1107,

p = 0.0560). The lack of correlation between GC4 skew and intronic GC skew appears to be explained, in part, by the C-enrichment near intron extremities. Indeed, after removing 1kb of sequence at both ends and only considering cores of introns > 2100bp, intronic and fourfold degenerate GC skews are positively correlated (rho = 0.0451, p = 0.0199, n = 2666) to a similar extent as AT skews (rho = 0.0490, p = 0.0114, n = 2666). That the relationship between GC4 skew and expression (rho = -0.1883, $p < 2.2 \times 10^{-16}$) is considerably stronger than seen for AT4 skew (rho= -0.0222, p = 0.0116, n = 12928) is consistent with selection on codon usage favouring C-ending codons (see Duret and Mouchiroud 1999), as well as the counteracting effect of ESE enrichment on AT skew.

Thus, because direct measures of male germline expression (rho = 0.3089, $p < 2.2 \times 10^{-16}$ for mitotic expression; rho = 0.2434, $p < 2.2 \times 10^{-16}$ for meiotic expression; rho = 0.3341, $p < 2.2 \times 10^{-16}$ for postmeiotic expression; n = 12543) and tau (rho = -0.3284, $p < 2.2 \times 10^{-16}$, n = 10901) correlate with somatic expression patterns (and hence with the strength of selection on codon usage) they are not suited to directly assessing the impact of germline transcription on compositional asymmetries in coding sequence. The above, however, suggests that selection for optimal codons or splicing efficiency have major influences on coding sequence.

Consistent with a role for selection, as opposed to mutation bias, we find that both RP (rho = 0.0543, $p = 6.231 \times 10^{-10}$, n = 12977) and ACE (rho = 0.0582, $p = 3.343 \times 10^{-11}$) are weakly positively correlated with fourfold degenerate AT skew (i.e. negatively with TA skew). This is also the case for GC4 skew and ACE (rho = 0.0302, p = 0.0006) but not RP (rho = 0.0139, p = 0.1123). Given the lack of a similar correlation for intronic skews, the most parsimonious explanation is that the above results reflect the activity

of Hill-Robertson interference modulating optimal codon usage, interference being diminished when recombination rates are high.

Discussion

Here, we have shown that intronic AT skew has features that make it a defendable proxy for germ-line transcription. Adding to the prior evidence (Touchon et al. 2004) that AT intronic skew in Drosophila is not influenced by proximity to intron extremities (unlike GC skew) and that it differs between transcribed DNA and flanking sequence, we also now establish that a) it is more similar than expected by chance between introns from the same gene (which is not accounted for by physical proximity), b) correlates with germ-line expression data, c) is more pronounced in broadly expressed genes. Furthermore, d) a trend for intronic skew to be different at 5' and 3' ends of genes is particular to broadly expressed genes. Finally, e) controlling for physical distance, introns of proximate genes are most different in skew if they have different tissue specificity. We do not argue that AT skew is solely owing to germline expression, but the above evidence is consistent with some component reflecting biases associated with transcription.

Given this measure, we see no evidence that germ-line expressed genes are prone to avoiding recombination nor that they evolve at unusual rates. The absence of correlation between AT skew and rates of protein evolution, however, masks an unexpected result, namely that for genes expressed during male meiosis, those with high rates of expression are fast evolving, while for genes abundantly expressed at other times in germline we observe the opposite correlation. Additionally, we find only very weak evidence that germline transcription determines skew at four-fold

degenerate sites via its effects on repair. Consequently, we cannot presently conclude that germ-line expression is likely to be a relevant factor in interpreting the correlation between rates of protein evolution and recombination (RP: rho = -0.0589, p = 8.88 $x10^{-8}$; ACE: rho = -0.0775, p = 1.854 $x10^{-12}$, n = 8243) observed in *D. melanogaster*. These negative results, nonetheless, come with the important caveat that while intronic AT skew reflects in part germ-line transcription it is not easily defendable that it is an especially strong predictor of germ-line transcription.

How can we relate our findings to the prior observation that broadly expressed genes, and hence those likely to be germline expressed, have low recombination rates (Larracuente et al. 2008)? One possibility is that the prior observation might be wrong. Indeed, we find that this result is sensitive to the measure of the recombination rate and subset of genes considered. When we restrict our analysis to genes with available divergence data, as in Larracuente et al. (2008), thus considering only those genes with well defined orthologs in six *Drosophila* species, we replicate the prior result and observe a significant positive correlation between RP and tau (rho = 0.0235, p = 0.0471, n = 7152). However, using ACE as the recombination measure, we see no effect (rho = -0.0029, p = 0.8046). If we employ a full gene set (not just those with defined orthologs) we now find that tau correlates negatively with ACE (rho = -0.0225, p = 0.0201, n = 10709), the opposite of the prior claim, but not RP (rho = 0.0134, p = 0.1643). These inconsistencies suggest that any relationship between recombination and tissue specificity is likely to be weak if present at all. The discrepancies between correlations with the two recombination measures we employed here, as well as the relatively low correspondence (r^2 = 0.475824) between the estimates for ACE provided by Larracuente et al. (2008) and Hey and Kliman (2002) imply a large degree of

uncertainty. It would be worthwhile to improve the resolution of recombination rate maps in *Drosophila*.

Aside from the above caveats, our results should be considered preliminary for other reasons. First, we cannot definitely rule out some coupling between recombination and germline expression immediately prior to female meiosis. Even if AT skew were a perfect as a measure of germ-line transcription, it would measure transcription in male and female germlines in all pre-meiotic cells from the zygote onwards. Much as the averaging effect of AT skew masks the difference in correlation between protein rates of evolution and levels of expression at different stages of germline development, if transcription immediately prior to meiosis were anti-recombinogenic (McVicker and Green 2010), our data would not necessarily reveal any such effect. Resolving this issue will require in-depth transcriptional and epigenetic profiles immediately prior to recombination in oocytes allied with high-resolution recombination maps.

In this light it is curious to note that RP (by necessity reflecting recombination in females) is weakly negatively correlated with premeiotic (rho = -0.0188, p = 0.0377) and meiotic (rho = -0.0213, p = 0.0181) expression in males (i.e. in testis), but not with postmeiotic testis expression (rho = -0.0143, p = 0.1129). While it is tempting to speculate that this may reflect a similarity in gene expression in testes and in oocytes, it should be noted that the effect is both very modest and not replicated when employing ACE as the recombination measure (ACE with premeiotic expression: rho = 0.0041, p = 0.653, n = 12251; meiotic expression: rho = -0.0037, p = 0.6818; postmeiotic expression: rho = -0.0061, p = 0.4987). Even the strongest predictor of this set has an r^2 of 0.05%. Moreover, the above evidence should not be considered as evidence that

germline expression drives recombination rates, even to a very limited degree. In centromeric regions and on chromosome IV, effects of transcription cannot drive any association that is observed between germline expression and crossing-over. When we remove genes with an RP of zero and consider only those genes with the potential to influence recombination, the correlations between RP and mitotic (rho = -0.0070, p = 0.4599, n = 11018) and meiotic testis expression (rho = -0.0118, p = 0.2165) disappear. This is not an effect of reduced sample size, as the probability of observing a correlation as extreme as, or more extreme than that observed in the full set of genes within subsamples of n = 11018 is 0.4934 for mitotic rates and 0.4879 for meiotic rates (based on 10000 replicates). Presumably, the loss of significance is owing to higher testis expression rates in regions with no recombination (Wilcoxon signed rank test difference for mitotic rates -0.2032, p = 0.0054; difference for meiotic rates -0.1834, p = 0.0122). We thus find no evidence for a direct mechanistic interaction between transcription in germ cells around the time of meiosis and crossover. We do, however, find evidence that transcription may be higher in domains with no recombination, as reported by Haddrill et al. (2008).

Overall then, the data fail to support a negative relationship between breadth of expression or putative germline transcription assayed by intronic AT skew and crossover rate. In mammals, by contrast, a robust relationship between tissue specificity and recombination is seen (Necsulea et al. 2009b). Similarly, genes expressed in germline have low recombination rates (McVicker and Green 2010). Given the association between methylation and recombination reported in humans (Necsulea et al. 2009b; Sigurdsson et al. 2009) and the absence of methylation in Drosophila, it may be that *D. melanogaster* differs systematically from mammals.

Acknowledgements

CCW is funded by the University of Bath. LDH is a Royal Society Wolfson Research Merit Award holder. We thank Amanda Larracuente for providing the recombination rate estimates.

Figure 1

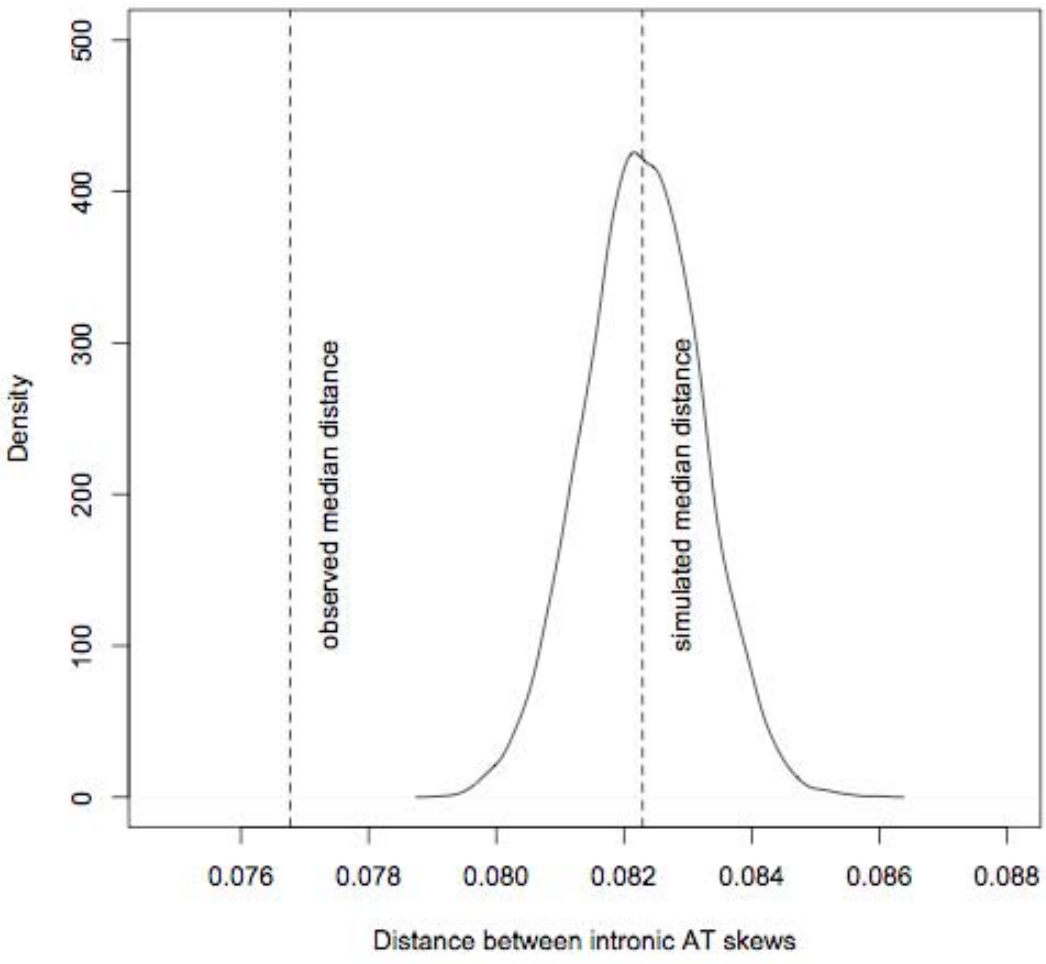

Figure 1 The median distance in AT skews between introns of the same gene is smaller than the median distance between randomly selected introns on the same chromosome. The curve represents the distribution of observed differences in skew.

Figure 2

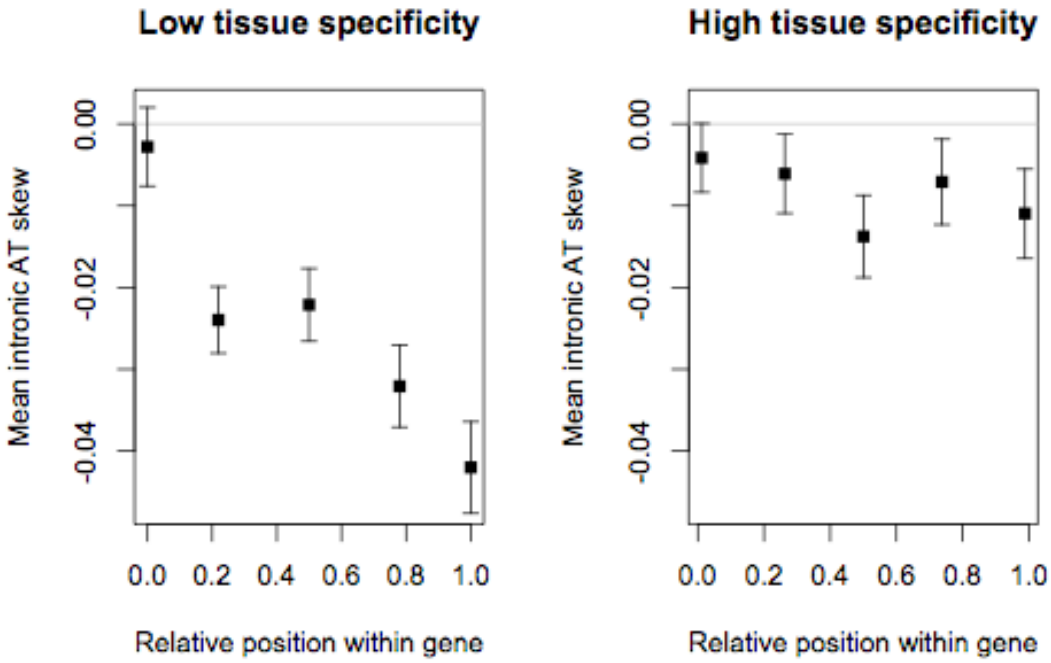

Figure 2 There is change in AT skew from 5' to 3' gene ends in low tissue specificity (defined as the lowest quartile for tau) but not high tissue specificity genes (defined as the highest quartile for tau). Genes were separated into 5 equal-sized bins based on their position relative to the most 5' (0) and the most 3' intron (1) and relative position for each bin was plotted against mean AT intron skew for that bin. Error bars represent the standard error of the mean.

Figure 3

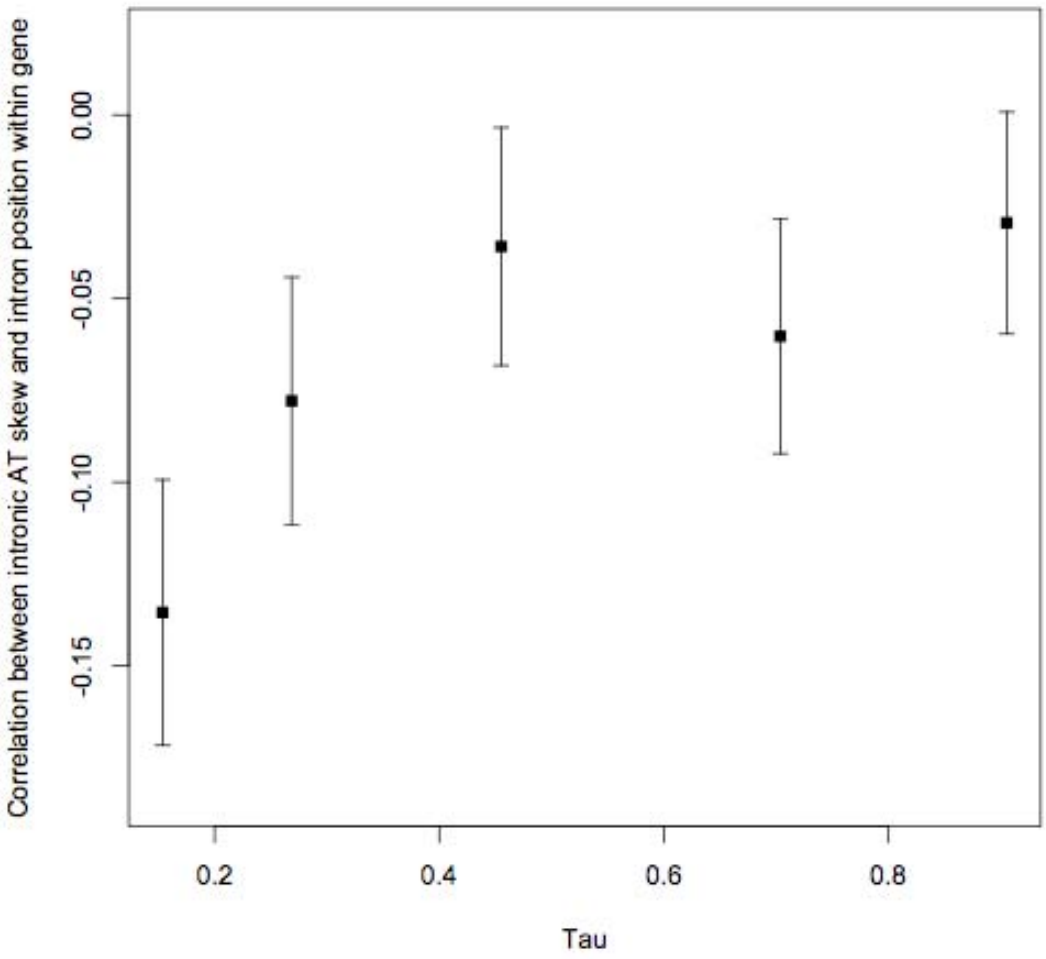

Figure 3 For broadly expressed genes (low tau), there is a stronger negative correlation between intronic AT skew within genes and intron position than for high-tau genes, indicating a more pronounced change in skew from 5’ to 3’ in broadly expressed genes. Genes were separated into 5 equal-sized bins by tau and mean tau for each bin was plotted against the mean correlation observed between intronic AT skew and relative 5’-3’ location. Error bars represent the standard error of the mean.

Figure 4

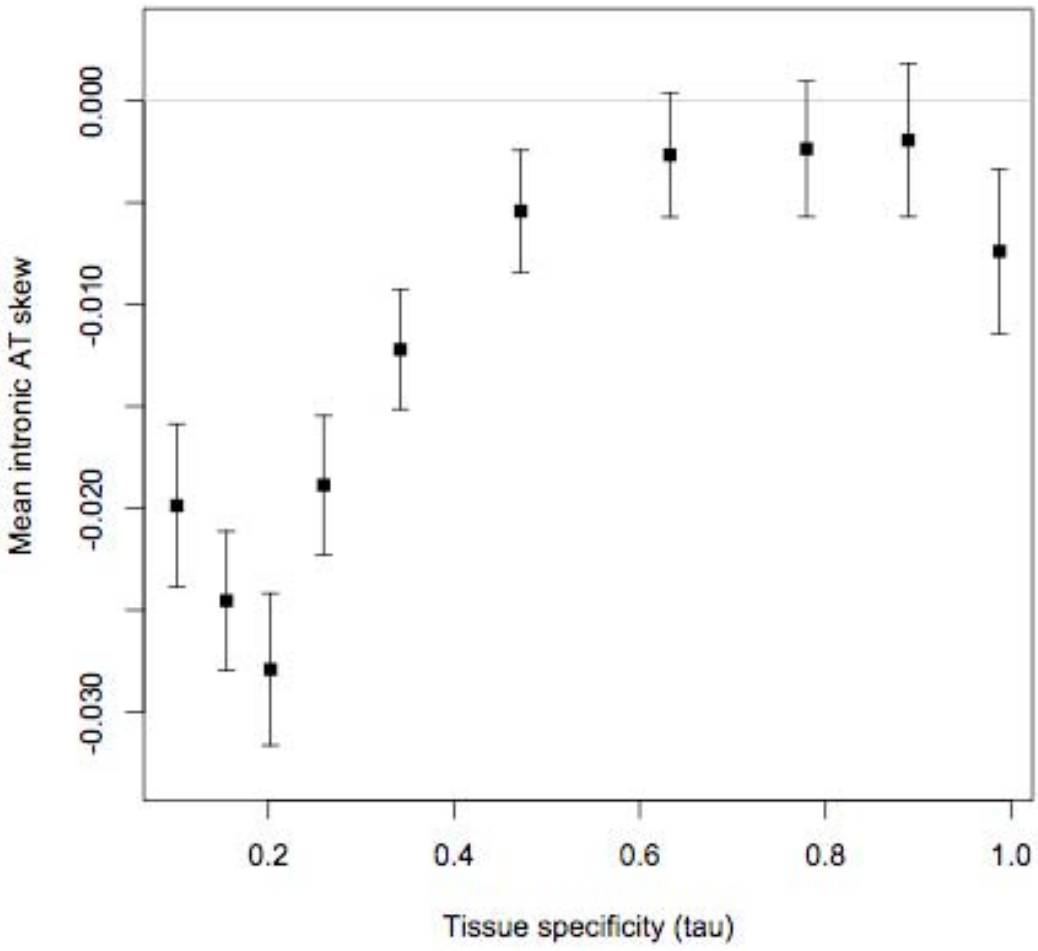

Figure 4 The magnitude of intronic AT skew decreases with increasing tissue specificity. Tissue specificity (tau) data were separated into 10 equal-sized bins and mean tau for each bin was plotted against the mean intronic AT skew for each bin. Error bars represent the standard error of the mean.

Figure 5

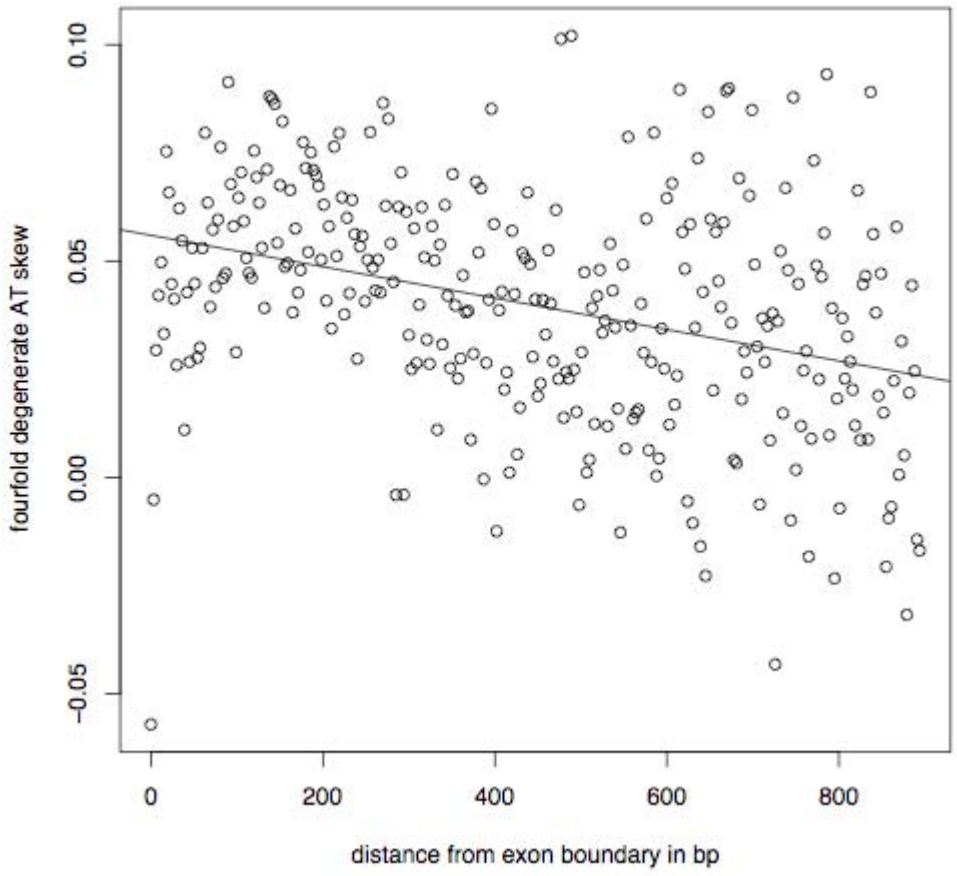

Figure 5 Fourfold degenerate AT skew decreases with distance from exon boundaries (rho = -0.3725, p = 3.935 x10^{-11}), indicating that the presence of A-rich exonic splice enhancers imposes a skew well into the coding sequence, hence making fourfold degenerate AT skew in coding sequences an unsuitable proxy for germline transcription.

References

Bachtrog D (2003) Protein evolution and codon usage bias on the neo-sex chromosomes of Drosophila miranda. Genetics 165:1221-1232

Berchowitz LE, Hanlon SE, Lieb JD, Copenhaver GP (2009) A positive but complex association between meiotic double-strand break hotspots and open chromatin in Saccharomyces cerevisiae. Genome Res 19:2245-2257

Berglund J, Pollard KS, Webster MT (2009) Hotspots of biased nucleotide substitutions in human genes. PLoS Biol 7:e26
Betancourt AJ, Presgraves DC (2002) Linkage limits the power of natural selection in Drosophila. Proc Natl Acad Sci USA 99:13616-13620
Betancourt AJ, Welch JJ, Charlesworth (2009) Reduced Effectiveness of Selection Caused by a Lack of Recombination. Curr Biol 19:655-660
Borde V, Robine N, Lin W, Bonfils S, Geli V, Nicolas A (2009) Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. EMBO J 28:99-111
Buard J, Barthes P, Grey C, de Massy B (2009) Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. EMBO J 28:2616-2624
Chintapalli VR, Wang J, Dow JA (2007) Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet 39:715-720
Comeron JM (2004) Selective and mutational patterns associated with gene expression in humans: Influences on synonymous composition and intron presence. Genetics 167:1293-1304
de Wit E, Braunschweig U, Greil F, Bussemaker HJ, van Steensel B (2008) Global chromatin domain organization of the Drosophila genome. PLoS Genet 4:e1000045
Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 10:285-311
Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, Arabidopsis. Proc Natl Acad Sci USA 96:4482-4487
Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol 17:68-74
Eyre-Walker A, Hurst LD (2001) The evolution of isochores. Nat Rev Genet 2:549-555
Galtier, Duret, Glemin, Ranwez (2009) GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet 25:1-5
Galtier N, Duret L (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet 23:273-277
Green P, Ewing B, Miller W, Thomas PJ, Program NCS, Green ED (2003) Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet 33:514-517
Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130:77-88
Haddrill PR, Charlesworth B (2008) Non-neutral processes drive the nucleotide composition of non-coding sequences in Drosophila. Biol Letters 4:438-441
Haddrill PR, Waldron FM, Charlesworth B (2008) Elevated levels of expression associated with regions of the Drosophila genome that lack crossing over. Biol Letters 4:758-761
Hey J, Kliman RM (2002) Interactions between natural selection, recombination and gene density in the genes of drosophila. Genetics 160:595-608
Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Arneodo A, Thermes C (2007) Human gene organization driven by the coordination of replication and transcription. Genome Res 17:1278-1285

Kulathinal RJ, Bennettt SM, Fitzpatrick CL, Noor MAF (2008) Fine-scale mapping of recombination rate in Drosophila refines its correlation to diversity and divergence. Proc Natl Acad Sci USA 105:10051-10056

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG (2008) Evolution of protein-coding genes in Drosophila. Trends Genet 24:114-123

Lercher MJ, Urrutia AO, Hurst LD (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet 31:180-183

Liao BY, Scott NM, Zhang JZ (2006) Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol 23:2072-2080

Majewski J (2003) Dependence of Mutational Asymmetry on Gene-Expression Levels in the Human Genome. Am J Hum Genet 73:688-692

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454:479-485

Marais G, Mouchiroud D, Duret L (2003) Neutral effect of recombination on base composition in Drosophila. Genet Res 81:79-87

McVicker G, Green P (2010) Genomic signatures of germline gene expression. Genome Res: doi: 10.1101/gr.106666.110; Epub date 2010/08/06

Mugal CF, Wolf JB, von Grunberg HH, Ellegren H (2010) Conservation of neutral substitution rate and substitutional asymmetries in mammalian genes. Genome Biol Evol 2:19-28

Nagylaki T (1983) Evolution of a finite population under gene conversion. Proceedings Of the National Academy Of Sciences Of the United States Of America-Biological Sciences 80:6278-6281

Necsulea A, Guillet C, Cadoret J-C, Prioleau M-N, Duret L (2009a) The Relationship between DNA Replication and Human Genome Organization. Mol Biol Evol 26:729-741

Necsulea A, Semon M, Duret L, Hurst LD (2009b) Monoallelic expression and tissue specificity are associated with high crossover rates. Trends Genet 25:519-522

Pal C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158:927-931

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD (2007) Splicing and the evolution of proteins in mammals. PLoS Biol 5:343-353

Polak P, Arndt PF (2008) Transcription induces strand-specific mutations at the 5' end of human genes. Genome Res 18:1216-1223

Powell JR, Moriyama EN (1997) Evolution of codon usage bias in *Drosophila*. Proc Natl Acad Sci USA 94:7784-7790

Presgraves DC (2005) Recombination enhances protein adaptation in Drosophila melanogaster. Curr Biol 15:1651-1656

Sigurdsson MI, Smith AV, Bjornsson HT, Jonsson JJ (2009) HapMap methylation-associated SNPs, markers of germline DNA methylation, positively correlate with regional levels of human meiotic recombination. Genome Res 19:581-589

Singh ND, Aquadro CF, Clark AG (2009) Estimation of Fine-Scale Recombination Intensity Variation in the white-echinus Interval of D-melanogaster. J Mol Evol 69:42-53

Spellman PT, Rubin GM (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol 1:5
Sun M, Hurst LD, Carmichael GG, Chen J (2005) Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts. Nucl. Acids Res. 33:5533-5543
Sun M, Hurst LD, Carmichael GG, Chen J (2006) Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcriptionand organismic complexity. Genome Res 16:922-933
Svejstrup JQ (2002) Mechanisms of transcription-coupled DNA repair. Nat Rev Mol Cell Biol 3:21-9
Takano-Shimizu T (2001) Local changes in GC/AT substitution biases and in crossover frequencies on Drosophila chromosomes. Mol Biol Evol 18:606-619
Touchon M, Arneodo A, d'Aubenton-Carafa Y, Thermes C (2004) Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. Nucleic Acids Res 32:4969-4978
Touchon M, Nicolay S, Audit B, Brodie of Brodie EB, d'Aubenton-Carafa Y, Arneodo A, Thermes C (2005) Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. Proc Natl Acad Sci USA 102:9836-9841
Vibranovski MD, Lopes HF, Karr TL, Long M (2009) Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. PLoS Genet 5:e1000731
Warnecke T, Hurst LD (2007) Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. Mol Biol Evol 24:2755-2762
Warnecke T, Parmley JL, Hurst LD (2008) Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. Genome Biol 9:R29
Weber CC, Hurst LD (2009) Protein Rates of Evolution Are Predicted by Double-Strand Break Events, Independent of Crossing-over Rates. Genome Biol Evol 1:340-349
Webster MT, Smith NG, Lercher MJ, Ellegren H (2004) Gene expression, synteny, and local similarity in human noncoding mutation rates. Mol Biol Evol 21:1820-1830
Wu TC, Lichten M (1994) Meiosis-induced double-strand break sites determined by yeast chromatin structure. Science 263:515-518
Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, Lancet D, Shmueli O (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650-659
Yang L, Yu J (2009) A comparative analysis of divergently-paired genes (DPGs) among Drosophila and vertebrate genomes. BMC Evol Biol 9:55
Zhang Y, Sturgill D, Parisi M, Kumar S, Oliver B (2007) Constraint and turnover in sex-biased gene expression in the genus Drosophila. Nature 450:233-237
Zhang Z, Hambuch TM, Parsch J (2004) Molecular evolution of sex-biased genes in Drosophila. Mol Biol Evol 21:2130-9

Zhang Z, Parsch J (2005) Positive correlation between evolutionary rate and recombination rate in Drosophila genes with male-biased expression. Mol Biol Evol 22:1945-1947